REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12/21/2011	Final Technical	1/31/2008 - 07/31/2011

4. TITLE AND SUBTITLE
Dynamic and Supervised Topic Models for Literature-Based Discovery

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-08-1-0487

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Blei, David

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Computer Science
35 Olden Street
Princeton University
Princeton, NJ 08544

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ONR Reg Boston N62879
495 Summer Street, Room 627
Boston, MA 02210-2109
617-753-3283 Phone
617-753-4605 Fax

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution is Unlimited

13. SUPPLEMENTARY NOTES
20120109030

14. ABSTRACT
Under the support of the ONR my research focused on extending the state of the art of probabilistic topic modeling, algorithms for making discoveries from and predictions about large collections of texts. For the past three years, my group has published many papers in the service of this goal. In this report, I will highlight some of the themes and publications that represent this work. Thanks to the support of the ONR, we have made excellent progress in our stated goals.

15. SUBJECT TERMS
bayesian computation, probabilistic topic modeling, massive data sets

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. David Blei, Professor
SAR	SAR	SAR	SAR	6	19b. TELEPONE NUMBER *(Include area code)* 609-258-9907

Dynamic and Supervised Topic Models for Literature-Based Discovery

Final Report to the Office of Naval Research

David M. Blei
Princeton University

Under the support of the ONR my research focused on extending the state of the art of *probabilistic topic modeling*, algorithms for making discoveries from and predictions about large collections of texts. For the past three years, my group has published many papers in the service of this goal. In this report, I will highlight some of the themes and publications that represent this work. Thanks to the support of the ONR, we have made excellent progress in our stated goals.

Bayesian nonparametric modeling

Topic models discover the latent themes that pervade a corpus of documents. Our broad goal is to build methods that can be applied widely, that is, to many kinds of corpora. This motivates our development of *Bayesian nonparametric (BNP) models*. BNP models adapt the cardinality of the discovered topics, such as the number of latent topics or the form of the topic hierarchy, to the corpus at hand. (In contrast, traditional parametric models require that the analyst specify this structure in advance.) We have developed new Bayesian nonparametric topic models and new scalable algorithms for BNP modeling.

- Sam Gershman and I wrote a tutorial about Bayesian nonparametrics [12]. There is a high bar to working with BNP models, as the literature has evolved from several fields. We hope that our tutorial will provide a clear introduction to the main ideas.

- We have explored several ways of building dependence into BNP models.

  Peter Frazier (Cornell) and I developed *distance dependent* Bayesian nonparametric models [1, 2, 13]. These allow external data sources to influence the latent clustering (and latent feature representation) of a variety of data. We used these models to capture sequential dependence in text and spatial dependence in images. We released open-source software that implements our algorithms.

In other work, Lauren Hannah (Duke) and Warren Powell and I developed *Dirichlet process mixtures of generalized linear models* [14, 15]. These allow covariates to affect the clustering of a response and exert a relationship on it. DP-GLMs allow us to fit appropriately complex response functions in prediction problems, fitting nonlinearity via several linear components.

- BNP topic models are hierarchical mixed-membership models of text, usually based on the hierarchical Dirichlet process. One thread of our research has been to build hierarchical BNP models that relax some of the limiting assumptions of the original HDP.

  John Paisley (Berkeley), Chong Wang, and I developed the *Discrete Infinite Logistic Normal* (DILN), which is a new kind of Bayesian nonparametric model [19]. (This paper won a **Notable Paper Award** at AI-STATS.) DILN allows the atoms of an underlying random measure to exhibit correlation. The DILN topic model is a BNP variant of the correlated topic model, allowing the appearance within a document of latent subjects (like health and sports) to be correlated. Unlike the correlated topic model, the number of topics is determined by the data.

  In other work, Chong Wang and I developed a hierarchical BNP topic model with "spike and slab" priors on the latent topics [20]. This gave better predictive performance, decoupling the sparsity of the topics and their smoothness, i.e., decoupling how many words a topic contains from how confident we are about their probabilities within it.

- Michael Jordan (Berkeley), Tom Griffiths (Berkeley), and I developed hierarchical latent Dirichlet allocation [3]. This is a BNP topic model that finds an arbitrary tree structure (of arbitrary depth) to describe the topics in a collection of documents. The prior distribution we developed for this model—the nested Chinese restaurant process—illustrates the advantages of BNP methods. While classical methods of model selection can be used to choose a simple number of components, these methods cannot help us search over the arbitrary space of tree structures.

  In more recent work, Chong Wang and I developed a fast variational inference algorithm for this model [24]. This is the first variational inference method that searches over combinatorial structures as part of the optimization.

- The research items above focus on mixture models or mixed-membership models. We have also worked on latent factor models, i.e., Bayesian nonparametric models of matrix factorization. Sinead Williamson (Carnegie-Mellon), Katherine Heller (Duke), Chong Wang, and I used BNP factor models with mixed-membership models to better control sparsity in determining how many topics are active in each document [25].

  In other work, John Paisley (Berkeley), Larry Carin (Duke), and I developed a fast variational inference algorithm for BNP factor models [18].

Dynamic topic models

We continue to research models of language that capture how language changes over time.

- Our earlier work on this subject assumed that time was discrete, for example we analyzed *Science* year by year. Chong Wang and I developed a continuous time dynamic topic model [22]. This lets documents appear with time-stamps at arbitrary granularity. Further, we can model language change at multiple resolutions.

- Sean Gerrish and I developed a dynamic topic model that captures *influential* documents [10]. Our model posits that an influential document is one that is prescient of how language changed. For example, Einstein's first paper about General Relativity was an influential paper because many papers discussed it subsequent to its publication. Our method infers the influence score of each document by analyzing a large corpus of sequentially ordered documents.

  To validate our method, we inferred influence scores on several large corpora of scientific articles and measured that our score correlates significantly to citation counts. I emphasize that our scores are only computed from the language of the articles themselves—our model could be used to find influential documents in corpora that do not contain citations. *The Economist* reported on this research ("Organising the Web: The Science of Science" April 28 2011).

Modeling networks and text

We have also developed new models of networks and their relationships to text.

- Jonathan Chang (Facebook) and I developed the *relational topic model*, which finds topics that respect the network connectivity of the documents [7, 6]. Unlike traditional network models, this model incorporates node content—it can predict content from links and links from content. Jonathan released open source software that implements his algorithm.

- Jonathan Chang (Facebook), Jordan Boyd-Graber (University of Maryland) and I developed a model that discovers the social network hidden inside texts [8]. The idea is to use named entities in the text and to identify when two named entities significantly co-occur. Further, we find patterns of words that describe these relationships. For example, we analyzed a corpus of New York Times articles to find related people in the news and to describe their relationships. The model discovered relationships described by by familial words, adversarial words, and others.

Models of text and other variables

Much of our recent work centers around using other kinds of variables to help anchor text models, and to use text models to predict other kinds of variables.

- Chong Wang and I developed a new method for collaborative filtering. Our method uses both user preferences and *content* about the items [21]. This work won the **Best Student Paper Award** at KDD 2011.

- Sean Gerrish and I built a model of legislative roll call data (i.e., votes on bills) and bill texts [11]. This extends classical quantitative political science models, which only model votes. This work won a **Distinguished Application Award** at ICML 2011. We are continuing to work on this area, building a new exploratory model of legislators that gives descriptions of how their votes deviate from otherwise typical patterns.

- Jordan Boyd-Graber (University of Maryland) and I have developed several methods for combining natural language processing data with topic models. In one project, we modeled multi-lingual corpora [4]. In another, we modeled constraints based on dependency parses and latent topics [5].

Current efforts

We have currently turned our attention to two important problems.

- First we are examining scalable computation for topic models. Matt Hoffman (Columbia), Francis Bach (INRIA), and I developed stochastic variational inference for Latent Dirichlet allocation [16]. This algorithm lets us analyze massive document collections, including document collections that arrive in a never-ending stream. Chong Wang and I extended this algorithm to the hierarchical Dirichlet process, enabling us to fit Bayesian nonparametric models to massive data [23].

- Second we are examining how we can better use topic models for interpretative and exploratory tasks, and examining how we might make this problem mathematically rigorous and well-defined. Jonathan Chang (Facebook), Jordan Boyd Graber (University of Maryland), Chong Wang, Sean Gerrish, and I implemented a large-scale user study with Amazon's Mechanical Turk to assess how interpretable topic models can be [9]. This was the first evaluation of unsupervised learning for interpretation with Mechanical Turk. (Since this paper, others have reproduced and emulated our experimental set-up.)

  In more recent work, David Mimno and I have explored *posterior predictive checks* for topic models [17]. This promises to be an automated way to assess which topics are interpretable, without needing to run a user-study for each fitted model.

References

[1] D. Blei and P. Frazier. Distance dependent Chinese restaurant processes. In *International Conference on Machine Learning*, 2010.

[2] D. Blei and P. Frazier. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461âĹŠ–2488, 2011.

[3] D. Blei, T. Griffiths, and M. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

[4] J. Boyd-Graber and D. Blei. Multilingual topic models for unaligned text. In *Uncertainty in Artificial Intelligence*, 2009.

[5] J. Boyd-Graber and D. Blei. Syntactic topic models. In *Neural Information Processing Systems*, 2009.

[6] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, 2009.

[7] J. Chang and D. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1), 2010.

[8] J. Chang, J. Boyd-Graber, and D. Blei. Connections between the lines: Augmenting social networks with text. In *Knowledge Discovery and Data Mining*, 2009.

[9] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 2009.

[10] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning*, 2010.

[11] S. Gerrish and D. Blei. Predicting legislative roll calls from text. In *International Conference on Machine Learning*, 2011.

[12] S. Gershman and D. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 2011.

[13] S. Ghosh, A. Ungureanu, E. Sudderth, and D. Blei. Spatial distance dependent Chinese restaurant processes for image segmentation. In *Neural Information Processing Systems*, 2011.

[14] L. Hannah, D. Blei, and W. Powell. Dirichlet process mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, 2010.

[15] L. Hannah, D. Blei, and W. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, to appear, 2011.

[16] M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2010.

[17] D. Mimno and D. Blei. Bayesian checking for topic models. In *Empirical Methods in Natural Language Processing*, 2011.

[18] J. Paisley, L. Carin, and D. Blei. Variational inference for stick-breaking beta processes. In *International Conference on Machine Learning*, 2011.

[19] J. Paisley, C. Wang, and D. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*, 2011.

[20] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Neural Information Processing Systems*, 2009.

[21] C. Wang and D. Blei. Collaborative topic moldeing for recommending scientific articles. In *Knowledge Discovery and Data Mining*, 2011.

[22] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence (UAI)*, 2008.

[23] C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2011.

[24] Chong Wang and David Blei. Variational inference for the nested Chinese restaurant process. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1990–1998. 2009.

[25] S. Williamson, C. Wang, K. Heller, and D. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *International Conference on Machine Learning*, 2010.